A NEW APPROACH TO SECOND ORDER COMPARISONS OF VARIANCE ESTIMATORS FOR COMPLEX SURVEYS

by

Cathy Campbell

Technical Report #343
March 1979

Department of Applied Statistics
University of Minnesota
St. Paul, Minnesota 55108

Abstract

Three methods of estimating the sampling variance of nonlinear statistics from complex surveys are reviewed--balanced repeated replication, jackknife replication and the Taylor methods. Other investigators have shown that von Mises expansion of a differentiable statistical function may be used to study these methods for independent, identically distributed observations. We show that the expansion can also be employed to approximate many important parameter estimates when the observations are obtained by sampling with replacement from a finite stratified population. Using this method, the bias of the jackknife replication variance estimator is evaluated to $O(L^{-2})$.

keywords: jackknife variance estimator
balance repeated replication
linearization
Taylor method
analytical surveys

1. Introduction

How to make reasonable inferences about nonlinear parameters of finite populations is a problem that often troubles survey statisticians. Exact properties of nonlinear statistics are not known even when simple random sampling procedures are used. When complex sample designs are employed, the problems become even more difficult.

Suppose we wish to form a 95% confidence interval for some nonlinear parameter θ, e.g., a correlation coefficient. The usual approach is to

(1) calculate $\hat{\theta}$, a point estimate of θ;

(2) calculate $\sqrt{\hat{V}(\hat{\theta})}$, an estimated standard error;

(3) <u>hope</u> that $T = (\hat{\theta}-\theta)/\sqrt{\hat{V}(\hat{\theta})}$ has an approximate t or N(0,1) distribution;

(4) let $\hat{\theta} \pm t_{.975}\sqrt{\hat{V}(\hat{\theta})}$ be the confidence interval.

Attempts to justify the above procedure have continued over many years. Except for some asymptotic results, the verification has been largely empirical. Particular emphasis has focused on variance estimation -- how to calculate $\hat{V}(\hat{\theta})$ so it is a "good" point estimator of $V(\hat{\theta})$. Less attention has been given to the distribution of T.

Three different methods of variance estimation -- balanced repeated replication, jackknife replication, and the Taylor method -- have emerged from essentially separate considerations. Numerical studies have shown important small sample differences; no single method appears to dominate for all situations and purposes.

Here we present a theoretical framework that encompasses all three methods and will allow small sample comparisons to be made among the methods for stratified-cluster samples. This is done by showing that the parameters and their estimates may be expressed as functionals of the appropriate distribution functions. Then the first two terms of the von Mises (1947) expansion of a differentiable statistical function can be used to approximate $\hat{\theta}$. Using this approximation straightforward calculations will lead to expectations and mean-squared errors of the variance estimators.

First we describe the sampling and estimation procedures and give a review of the three methods of variance estimation. Use of the von Mises expansion for iid random variables is discussed before continuing to the case of stratified sampling. Finally some preliminary results are given.

2. Sample Design and Parameter Estimation

The sample design considered in this study is a useful simplification of designs used by the Bureau of the Census and other organizations which conduct area samples of human populations. The design can be characterized as a stratified-cluster sample with two units (clusters) chosen randomly with replacement from each stratum. To make the description more precise, we introduce the following notation:

L = number of strata in the population

A_i = number of clusters in stratum i $(i=1, \ldots, L)$

n_{ij} = number of elements in cluster j of stratum i

$(i = 1, \ldots L; j = 1, \ldots, A_i)$.

For the purposes of this study we presume that a bivariate data vector is associated with each element in the population. We represent the data as

$$\{(X_{ijk}, Y_{ijk}): \ i = 1, \ldots, L; \ j = 1, \ldots A_i; \ k = 1, \ldots n_{ij}\}.$$

For each cluster we define the elements of a 6×1 data vector $\underline{U}_{ij}$ as

$$U_{ij1} = \sum_{k=1}^{n_{ij}} X_{ijk}$$

$$U_{ij2} = \sum_{k=1}^{n_{ij}} Y_{ijk}$$

$$U_{ij3} = \sum_{k=1}^{n_{ij}} X_{ijk}^2$$

$$U_{ij4} = \sum_{k=1}^{n_{ij}} Y_{ijk}^2$$

$$U_{ij5} = \sum_{k=1}^{n_{ij}} X_{ijk}Y_{ijk}$$

$$U_{ij6} = n_{ij} \ .$$

The parameters of interest here are functions of the vector of population totals

$$\underline{U} = \sum_{i=1}^{L} \sum_{j=1}^{A_i} \underline{U}_{ij} = (U_1, U_2, U_3, U_4, U_5, U_6)' \quad .$$

For example

$$\overline{Y} = \frac{U_2}{U_6}, \qquad \sigma_x^{\ 2} = \frac{U_3}{U_6} - \left(\frac{U_1}{U_6}\right)^2 ,$$

and

$$B_{y \cdot x} = \frac{U_5 - U_1 U_2 / U_6}{U_3 - U_1^{\ 2} / U_6} \quad .$$

The usual estimates for parameters of this class are formed by substituting unbiased estimates of the U_m into the appropriate defining expression.

With two independent selections per stratum, the sample data are $\{\underline{u}_{ij}: \ i = 1, \ . \ . \ . \ , L; \ j = 1,2\}$. Then

$$\hat{\underline{U}} = \sum_{i=1}^{L} \frac{A_i}{2} (\underline{u}_{i1} + \underline{u}_{i2}) \tag{2.1}$$

is the appropriate unbiased estimator of $\underline{U}$. The estimator of $\sigma_x^{\ 2}$ is, for example,

$$\hat{\sigma}_x^{\ 2} = \frac{\hat{U}_3}{\hat{U}_6} \quad \left(\frac{\hat{U}_1}{\hat{U}_6}\right)^2 .$$

Other estimators are defined in a similar manner. In the next section we describe the currently used methods of estimating the sampling variance of these nonlinear statistics.

3. Methods of Variance Estimation

Unfortunately the standard theory for finite population sampling does not provide general methods for estimating the sampling variance of nonlinear statistics. In the absence of exact theory, three general methods of variance estimation for nonlinear statistics have evolved. While theoretical properties of these procedures are not known, they have developed because they made sense, reduced to known results for linear statistics, and have performed well in numerical studies. After describing the three methods, we shall briefly review the theoretical and numerical results that support their use.

3.1 Balanced Repeated Replication

Balanced repeated replication (BRR) or half-sample replication apparently arose in an attempt to approximate the variance estimate that can be obtained if independent replicated samples are available. McCarthy (1966, 1969a, 1969b), Kish and Frankel (1970), and Frankel (1971), among others, have published descriptions of the method.

For general exposition let $\theta = f(\underline{U})$ be the scalar-valued parameter of interest, and let $\hat{\theta} = f(\hat{\underline{U}})$ be the estimator using all of the sample data. A half-sample consists of one unit (either $\underline{u}_{i1}$ or $\underline{u}_{i2}$) from each stratum (i=1, . . . , L). The estimate of $\underline{U}$ based on the ℓ^{th} half-sample is

$$\hat{\underline{U}}^B_\ell = \hat{\underline{U}} + \sum_{i=1}^{L} \frac{A_i}{2} m_{\ell i} (\underline{u}_{i1} - \underline{u}_{i2})$$

where

$$m_{\ell i} = \begin{cases} +1 \text{ if } \underline{u}_{i1} \text{ is chosen in } \ell^{th} \text{ half-sample} \\ -1 \text{ if } \underline{u}_{i2} \text{ is chosen in } \ell^{th} \text{ half-sample} . \end{cases}$$

The ℓ^{th} half-sample estimate of θ is

$$\hat{\theta}_\ell^B = f(\underline{\hat{U}}_\ell^B) .$$

If r different half-samples are formed, then

$$\hat{V}_B(\hat{\theta}) = \frac{1}{r} \sum_{\ell=1}^{r} (\hat{\theta}_\ell^B - \hat{\theta})^2 \tag{3.1}$$

is the BRR estimator of $V(\hat{\theta})$.

A set of half-samples is balanced if $\hat{V}_B(\hat{\theta})$ reduces to the usual variance estimator when $\hat{\theta}$ is linear. Balanced sets are formed by choosing the $m_{\ell i}$ so that the rxL matrix M satisfies

a) $m_{\ell i} = +1$ or -1

b) $M'M = rI_L$

c) $M'\underline{1}_r = \underline{0}$.

The smallest r for which a balanced set can be found is the first multiple of 4 greater than L. The matrices M are subsets of the columns of rxr Hadamard matrices which are design matrices for 2^L fractional factorial designs (Plackett and Burman, 1948).

Each half-sample defines a complement half-sample which contains the "other" unit from each stratum. Some variations of the BRR variance estimate also use information from the complement half-samples. See Frankel (1971). Another version is formed by using $\overline{\theta}^B$, the average of the $\hat{\theta}_\ell^B$, instead of $\hat{\theta}$ in (3.1).

3.2 Jackknife Replication

The sample survey version of the jackknife variance estimator was first described by McCarthy (1966). Up to this point it appears that the jackknife had been used only with identically distributed data; in this case the "drop out 1" point estimate of θ (called a pseudoestimate here) is unambiguously defined. With stratified sampling the functional form of the pseudoestimates has been changed so that each pseudoestimate of $\underline{U}$ is unbiased.

Let $\hat{\theta}^J_{ij}$ be the jackknife pseudoestimate formed by deleting the $(i,j)^{th}$ sample unit. Then $\hat{\theta}^J_{ij} = f(\underline{\hat{U}}^J_{ij})$ where

$$\underline{\hat{U}}^J_{ij} = A_i \, \underline{u}_{ij'} + \sum_{k\neq i}^{L} \frac{A_k}{2} (\underline{u}_{k1} + \underline{u}_{k2})$$

with $\underline{u}_{ij'}$ being the undeleted observation in stratum i.

Several forms of the jackknife variance estimator have appeared in the literature (Frankel, 1971; Hislop and Lemeshow, 1977). The most common jackknife variance estimator is based on L pseudoestimates formed by randomly choosing one observation from the i^{th} stratum to be deleted for the i^{th} pseudoestimate. The full jackknife variance, based on 2L pseudoestimates, is given by

$$\hat{V}_J(\hat{\theta}) = \frac{1}{2} \sum_{i=1}^{L} \sum_{j=1}^{2} (\hat{\theta}^J_{ij} - \hat{\theta})^2 . \tag{3.2}$$

Alternatively, $\overline{\theta}^J$, the average of the pseudoestimates, can be used instead of $\hat{\theta}$ in (3.2).

3.3 Taylor Approximation

The third variance estimator often considered for complex surveys has a long history in statistics and in sampling. Known as the δ-method, Taylor method, propogation of errors, or the linearization method, this technique is used in all sampling texts to derive an approximation to the variance of the ratio of two sample sums. Tepping (1968) was the first to apply the Taylor method to general nonlinear statistics for complex surveys. The form of the estimator is given by

$$\hat{V}_T(\hat{\theta}) = \sum_{k=1}^{6} \left(\frac{\partial f(\underline{U})}{\partial U_k} \right)^2_{\underline{U}=\hat{\underline{U}}} \hat{V}(\hat{U}_k) + 2 \sum_{k<j}\sum \left(\frac{\partial f(\underline{U})}{\partial U_k} \cdot \frac{\partial f(\underline{U})}{\partial U_j} \right)_{\underline{U}=\hat{\underline{U}}} \hat{\text{cov}}\ (\hat{U}_k, \hat{U}_j) \ . \tag{3.3}$$

Since the elements of $\hat{\underline{U}}$ are linear functions of the sample data, the variances and covariances in (3.3) can be obtained using standard methods.

3.4 Properties of the Methods

Justification for using these methods of variance estimation for complex surveys comes from three sources:

(a) asymptotic behavior

(b) behavior when $\hat{\theta}$ is linear

(c) numerical studies.

The asymptotic behavior is what we would like, but does not provide a basis for choosing among the procedures. Krewski and Rao (1978) have shown that

$$V_I(\hat{\theta}) \xrightarrow{P} V(\hat{\theta}) \text{ as } L \longrightarrow \infty \ , \tag{3.4}$$

and

$$\frac{\hat{\theta} - \theta}{\sqrt{\hat{V}_I(\hat{\theta})}} \xrightarrow{\mathcal{L}} N(0,1) \text{ as } L \longrightarrow \infty \ , \tag{3.5}$$

where I = B,J, or T. These results lend some theoretical support to the common procedure of using

$$\hat{\theta} \pm z_{1-\alpha/2} \sqrt{\hat{V}_I(\hat{\theta})}$$

as a $100(1 - \alpha)\%$ confidence interval for θ.

Mellor (1973) studied a general class of balanced "drop out m" replication variance estimators for data from simple random samples. Besides showing that (3.4) and (3.5) hold as $n \longrightarrow \infty$, he also derived some interesting results when $\hat{\theta}$ is a sample mean or total. Within this setting he found that the MSE of the variance estimator is minimized for the "drop out 1" (ordinary jackknife) variance estimator. He also showed that the replication form of the t-statistic coincides with Walsh's (1947) t-statistic for correlated observations when $r = n$. McCarthy (1969) had also obtained this coincidence, under severe assumptions, when investigating BRR for stratified sampling. No analagous results have been obtained for nonlinear statistics, however.

To compensate for the lack of theoretical results and to investigate small sample properties, several numerical studies have been conducted -- some with real survey data, others with computer-generated data. The results show some consistency but have not led to a "best" procedure; nor have they provided a general understanding of the methods.

Frankel (1971) used data from the Current Population Survey in his large study which compared $\hat{V}_B$, $\hat{V}_J$, and $\hat{V}_T$ for 48 different parameters with $L = 6$, 12, and 30. Mellor (1973) used artificial data from simple random samples to compare V_T and "drop out m" procedures for $n = 16$ and 31. Campbell and Meyer (1978) generated nine stratified-clustered populations with $L = 6$ or 12 and compared $\hat{V}_B$, $\hat{V}_J$, and $\hat{V}_T$ for 12 different parameters. All investigations have shown that $\hat{V}_T$ produces consistently poorer confidence intervals than the replication methods. Frankel, and Campbell and Meyer found that versions of $\hat{V}_B$ gave the best confidence intervals, but even the best was not always adequate at these sample sizes. Confidence intervals were most reliable for ratios and regression coefficients, and worst for variances and correlations. At the same time both Frankel and Mellor found that $\hat{V}_T$ had the smallest relative MSE as an estimator of $V(\hat{\theta})$ even though, in Mellor's work, it suffered from a relative bias of about -30%. Mellor's replication methods gave variance estimates with large positive bias.

While the numerical studies have shown that real small sample differences between the methods do exist, they do not explain why the procedures perform differently. Sub-asymptotic results are needed for this. In the next section we present an approximation that allows $O(n^{-2})$ comparisons of the methods with simple random samples from infinite populations. Following that, we show that the same approximation can be used for stratified-cluster samples from finite populations.

4. von Mises Expansion for IID Observations

Hinkley (1978) and Jaeckel (1972) have shown that the von Mises (1947) expansion of a statistical differentiable function may be fruitfully employed to study the jackknife and Taylor variance estimators for iid data. Most of the material in this section is from their work. We shall describe the expansion and its relationship to the variance estimators under consideration. Mathematical rigor will be sacrificed here in the interests of clarity and brevity. Interested readers may consult Reeds (1977) for more detail. Suppose that

(a) $y_1, \ldots, y_n$ (possibly vectors) are iid observations from distribution F;

(b) $\theta = t(F)$ is a differentiable functional of F;

(c) $\hat{\theta} = t(\hat{F})$ where $\hat{F}$ is the sample cumulative distribution function.

Then

$$\hat{\theta} - \theta = \frac{1}{n} \sum_{i=1}^{n} t_1(y_i) + \frac{1}{2n^2} \sum_{ij}^{nn} t_2(y_i, y_j) + \ldots$$

$$+ \frac{1}{k!n^k} \sum\sum^{nn} \cdots \sum^{n} t_k(y_{i_1}, \ldots, y_{i_k}) + \ldots \quad . \quad (4.1)$$

The functions $t_1(y), t_2(y,z), \ldots$ are the 1^{st}, 2^{nd}, ... derivatives of the functional t(G) evaluated at G = F.

We shall assume that $\hat{\theta}-\theta$ can be approximated by the first two terms of the expansion. The derivatives $t_1(y)$ and $t_2(y,z)$ are found by solving

$$\frac{d}{d\varepsilon}\left[t\left((1-\varepsilon)F + \varepsilon G\right)\right]_{\varepsilon=0} = \int t_1(y)\, dG(y), \quad (4.2)$$

and

$$\frac{d^2}{d\varepsilon^2}\left[t\left((1-\varepsilon)F+\varepsilon G\right)\right]_{\varepsilon=0} = \int t_2(y,z)dG(y)dG(z) \ , \qquad (4.3)$$

where G is an arbitrary distribution function, subject to the constraints

$$E_F[t_1(y)] = 0$$

$$E_F[t_2(a,y)] = 0 \quad \text{for } a = \text{any constant}$$

$$t_2(y,z) = t_2(z,y) \quad .$$

The function $t_1(y)$ is the influence function of $\hat{\theta}$ which has played an important role in the theory of robust estimation (Hampel, 1974).

When n is large enough

$$\hat{\theta} - \theta \sim \frac{1}{n}\sum_{i=1}^{n} t_1(y_i) \qquad (4.4)$$

in which case the asymptotic behavior of $\hat{\theta}$ is obvious. To estimate $V(\hat{\theta})$ it is necessary to estimate $V[t_1(y)]$. The Taylor method and jackknife approach this in different ways. When (4.4) is valid, the i^{th} jackknife pseudovalue is simply

$$P_i = n\hat{\theta} - (n-1)\hat{\theta}_i = \theta + t_1(y_i)$$

where $\hat{\theta}$ is the pseudoestimate obtained by deleting y_i. The sample variance of the pseudovalues is an estimator of $V[t_1(y)]$. If the (4.4) is not valid, P_i will also depend on higher order terms in the expansion.

The Taylor estimate of $V(\hat{\theta})$ is also an approximation of $V[t_1(y)]$. We note that $t_1(y)$ depends on the parent distribution F; we make this explicit by writing $t_1(y) = t_1(y;F)$. The Taylor variance estimate, then, is

$$\hat{V}_T(\hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \hat{t}_1^2(y_i)$$

where $\hat{t}_1(y) = t_1(y;\hat{F})$. A simple example occurs when $\theta = \mu_y/\mu_x$ and $\hat{\theta} = \overline{y}/\overline{x}$. In this case

$$t_1(x,y) = \frac{1}{\mu_x}(y - \theta x) \quad ,$$

$$\hat{t}_1(x,y) = \frac{1}{\bar{x}}(y-\hat{\theta}x)$$

and

$$\hat{V}_T(\hat{\theta}) = \frac{1}{n\bar{x}^2} \sum_{i=1}^{n} (y_i - \hat{\theta}x_i)^2$$

which is the well-known first order approximation for the variance of a ratio.

Jaeckel, and Gray, Schucany and Watkins (1975) have developed $\hat{V}_T$ as a limiting case of the jackknife estimate as the disturbance on the deleted observation approaches 0.

Hinkley used the first two terms of the von Mises expansion to show that to $0(n^{-2})$

$$nV(\hat{\theta}) = \sigma_{11} + \frac{1}{n}(\sigma_{12} + \frac{1}{2}\sigma_{22}) \tag{4.5}$$

and $$E[\hat{V}(P_i)] = \sigma_{11} + \frac{1}{n}(\sigma_{12} + \sigma_{22}) \tag{4.6}$$

where $$\sigma_{11} = V[t_1(y)]$$

$$\sigma_{12} = \text{cov}\ [t_1(y),\ t_2(y,y)]$$

$$\sigma_{22} = V[t_2(y,z)]$$

$$\hat{V}(P_i) = \frac{1}{n-1} \sum_{i=1}^{n} (P_i - \bar{P})^2 \quad .$$

Simple results such as those given in (4.5) and (4.6) help explain the positive bias of $\hat{V}_J(\hat{\theta})$ found by Mellor: $E[\hat{V}(P_i)] - nV(\hat{\theta}) = \frac{1}{2}\sigma_{22}$ with $\sigma_{22} > 0$. The Taylor estimator $\hat{V}_T(\hat{\theta})$ will not, by its nature, account for the second order terms; unless $\sigma_{12} + \frac{1}{2}\sigma_{22} < 0$, $\hat{V}_T(\hat{\theta})$ will tend to underestimate $V(\hat{\theta})$. This phenomenon was also observed by Mellor.

Calculations, such as the above, can be carried out for other versions of jackknife and subsample variance estimators. The von Mises expansion is uniquely suited to the study of replication variance estimators because it identifies the contribution of each sample datum. Approximations for subsample estimates can be easily obtained from (4.1).

We now show that the advantages of the von Mises expansion can also be used to study variance estimation for stratified-cluster samples.

5. von Mises Expansion for Stratified Samples

The preceding results for simple random sampling show that $\hat{V}_B$, $\hat{V}_J$, and $\hat{V}_T$ are in a class of estimators whose properties may be explored via the von Mises expansion. In this section we show that the same approach can be used to study estimation for stratified sampling. This approach provides a sorely needed theoretical framework for comparing the variance estimators. The tool we use has been available since von Mises' work appeared in 1947, but its application to finite population sampling has not been recognized until now.

For stratified sampling with replacement, the data $\{\underline{u}_{ij}$: $i = 1, \ldots, L;$ $j = 1,2\}$ are realizations of independent, but not identically distributed, random variables. Although Reeds (1977), Filippova (1962), and other proponents of the von Mises method have been concerned with iid observations, von Mises original results were derived for the more general setting of arbitrary collections of independent random variables. As a result it is not necessary to prove new theorems, but merely to show that the parameters and their estimates may be expressed as functionals of the appropriate distribution functions.

For the exposition here, we work with the transformed data

$$\underline{W}_{ij} = A_i \, \underline{U}_{ij} \quad . \tag{5.1}$$

This is done so that the average $\frac{1}{A_i} \sum_{j=1}^{A_i} \underline{W}_{ij}$ of the data within a stratum is equal to $\underline{U}_{i.}$, the vector of totals for stratum i. Let $F_i(\underline{w})$ be the distribution function of the $\underline{W}_{ij}$ for the finite population of clusters in stratum i. With two independent selections per stratum, we obtain a realization of $\{\underline{W}_{ij}$: $i=1, \ldots, L;$ $j=1,2\}$ with $\underline{W}_{ij} \sim F_i(\underline{w})$. The average distribution function is

$$\overline{F}(\underline{w}) = \frac{1}{L} \sum_{i=1}^{L} F_i(\underline{w}).$$

Let $S(\underline{w})$ be the sample distribution function defined in the ordinary manner, as if the random variables were identically distributed.

A restatement of von Mises theorem says: if

(a) $\theta = g(\overline{F})$ is a differentiable functional of $\overline{F}$

(b) $\hat{\theta} = g(S)$ is the estimator of θ, then

$$\hat{\theta} - \theta \sim \overline{t}_1 = \frac{1}{2L} \sum_{i=1}^{L} \sum_{j=1}^{2} t_1(\underline{w}_{ij}) \quad \text{as } L \longrightarrow \infty \quad .$$

The first derivative $t_1(\underline{w})$ is the influence function evaluated at $\overline{F}$. Under appropriate restrictions on the moments of $\{t_1(\underline{W}_{ij})\}$, the Liapunov central limit theorem (Rao 1973, p. 127) gives

$$\sqrt{2L}\, \overline{t}_1 \xrightarrow{\mathcal{L}} N\left(0, \frac{1}{L} \sum_{i=1}^{L} V_1(i)\right) \quad \text{as } L \longrightarrow \infty \quad ,$$

where $\quad V_1(i) = V_{F_i}[t_1(\underline{W})] \quad .$

We now show that conditions (a) and (b) are satisfied for parameters which depend on ratios of the elements of $\underline{U}$. This restriction is necessary to insure that neither the parameter nor its estimate depends explicitly on the sample size 2L. This class of parameters includes means, proportions, variances, ratios of means, regression coefficients, and correlation coefficients, among others. These are the parameters for which the variance estimation methods are most commonly used. Let W_k be the k^{th} element of a general vector $\underline{W}$. By using (5.1) and comparing with (2.1), we observe that

$$\int w_k dS(\underline{w}) = \frac{1}{2L} \sum_{i=1}^{L} \sum_{j=1}^{2} (\underline{w}_{ij})_k = \frac{1}{L} \hat{U}_k \quad .$$

Similarly

$$\int w_k \, d\bar{F}\,(\underline{w}) = \frac{1}{L} \sum_{i=1}^{L} \int w_k dF_i(\underline{w}) = \frac{1}{L} U_k \quad .$$

As a result the ratios that appear in the parameter definitions,

$$\frac{U_i}{U_j} = \frac{\int w_i d\bar{F}(\underline{w})}{\int w_j d\bar{F}(\underline{w})} \quad ,$$

are functionals of $\bar{F}$, and the corresponding estimators are functionals of S.

Our goal in establishing the validity of von Mises results for finite population sampling is not to prove asymptotic results; these have been established using other methods. Instead we shall approximate $\hat{\theta}$ with the first two terms of the expansion in order to study the behavior of the variance estimators for finite L. The second order approximation for a general $\hat{\theta}$ is given by

$$\hat{\theta} - \theta \sim \frac{1}{2L} \sum_{i=1}^{L} \sum_{j=1}^{2} t_1(\underline{w}_{ij}) + \frac{1}{2(2L)^2} \sum_{i,j} \sum_{k,\ell} t_2\,(\underline{w}_{ij}, \underline{w}_{k\ell}) \qquad (5.2)$$

where $t_1(\underline{w}) = t_1(\underline{w};\bar{F})$ and $t_2(\underline{w},\underline{z}) = t_2(\underline{w},\underline{z};\ \bar{F})$ as defined in (4.2) and (4.3).

6. Examples and Conclusion

By applying the von Mises expansion to finite population sampling, we have found an approximation for $\hat{\theta}$ that should allow second order comparisons of three common methods of variance estimation. No assumptions have been made other than independent selections within each stratum. In particular, it is not necessary to assume a super-population model, although the method can be used in a super-population setting.

Comparing the variance estimators will be tedious; much of the work must be done numerically, but we now have a framework in which to work. At this time only limited results are available. Using (5.2) and some perseverance we have shown that to $O(L^{-2})$

$$\begin{aligned} Var(\hat{\theta}) = \frac{1}{2L^2} \sum_{i=1}^{L} V_1(i) + \frac{1}{8L^4} \sum_{i\neq j}^{L}\sum^{L} V_3(i,j) - \frac{1}{4L^4} \sum_{i\neq j}^{L}\sum^{L} C_4(i;j,j) \\ + \frac{1}{4L^3} \Sigma C_1(i) - \frac{1}{2L^3} \Sigma C_2(i,i) \end{aligned} \qquad (6.1)$$

where $V_1(i) = V_{F_i} [t_1(\underline{W})]$

$V_3(i,j) = V_{F_i,F_j} [t_2(\underline{W}, \underline{Z})]$

$C_4(i;j,j) = Cov_{F_i,F_j} [t_2(\underline{W},\underline{Z}), t_2(\underline{W},\underline{Y})]$

where $\underline{W} \sim F_i$, $\underline{Z}, \underline{Y} \sim F_j$

$C_1(i) = Cov_{F_i} [t_1(\underline{W}), t_2(\underline{W},\underline{W})]$

$C_2(i,i) = Cov_{F_i,F_i} [t_1(\underline{W}), t_2(\underline{W},\underline{Z})]$.

Let R be defined by $Var(\hat{\theta}) = \frac{1}{2L^2} \sum_{i=1}^{L} V_1(i) + R$. From (5.2) we can also find an approximation for a jackknife pseudoestimate as

$$\hat{\theta}_{p1} - \theta \sim \frac{1}{2L}\{\sum_{i\neq p}^{L}\sum_{j=1}^{2} t_1(\underline{w}_{ij}) + 2t_1(\underline{w}_{p2})\}$$

$$+ \frac{1}{2(2L)^2}\sum_{\substack{i,j\\ i,k\neq p}}\sum_{k,\ell} t_2(\underline{w}_{ij}, \underline{w}_{k\ell}) + 4\sum_{k\neq p}^{L}\sum_{\ell=1}^{2} t_2(\underline{w}_{p2}, \underline{w}_{k\ell})$$

$$+ 4t_2(\underline{w}_{p2}, \underline{w}_{p2}) \quad . \tag{6.2}$$

Then we can show that $E[\hat{V}_J(\hat{\theta})]$ as given in (3.2) is

$$E[\hat{V}_J(\hat{\theta})] = \frac{1}{2L^2}\sum_{i=1}^{L} V_1(i) + 2R \; . \tag{6.3}$$

(Details of (6.1) and (6.3) are given in the Appendix.)

We see that $\hat{V}_J(\hat{\theta})$ is positively or negatively biased depending on whether R is positive or negative. The mean-squared error of $\hat{\theta}$ is

$$MSE(\hat{\theta}) = V(\hat{\theta}) + B^2(\hat{\theta})$$

with $B(\hat{\theta}) = E[\hat{\theta}-\theta] = \frac{1}{2L^2}\sum_{i=1}^{L}[E_2(i) - E_3(i,i)]$

where
$$E_2(i) = E_{F_i}[t_2(\underline{W},\underline{W})]$$
$$E_3(i,i) = E_{F_i,F_i}[t_2(\underline{W},\underline{Z})] \quad .$$

As a result

$$E[\hat{V}_J(\hat{\theta})] - MSE(\hat{\theta}) = R - B^2(\hat{\theta}) \quad . \tag{6.4}$$

Expressions such as (6.4) will have to be evaluated numerically for specific estimators and distributions, but this should be more economical and enlightening than the large-scale simulations that have been done in the past.

Work has begun in a number of areas and will continue. One goal is a result like (6.3) for $E[\hat{V}_B(\hat{\theta})]$. We are obtaining analytical expressions for the terms in (6.1) for $\theta = U_1/U_2$ and $\theta = \sigma_x^2$. The more important task of evaluating the MSE of the variance estimates will also be undertaken.

We have reviewed the primary methods of estimating the variance of nonlinear statistics for complex surveys. A theoretical structure is given that will allow exact second order calculations of the bias and MSE of the variance estimators. Some results are given for the jackknife variance estimator. Work is continuing on other calculations.

Acknowledgment: This work was supported in part by Joint Statistical Agreement 78-79 with the Bureau of the Census, U.S. Department of Commerce.

<u>References</u>

Campbell, C. and M. M. Meyer (1978). Some properties of t-confidence intervals for survey data. <u>Proceedings of the Survey Research Section</u>, to appear. American Statistical Association, Washington, D.C.

Filippova, A. A. (1962). Mises' theorem on the asymptotic behavior of functionals of empirical distribution functions and its statistical applications. <u>Theory of Probability and its Applications</u>, <u>7</u>, 24-57.

Frankel, M. R. (1971). <u>Inference from Survey Samples: An Empirical Investigation</u>. The University of Michigan, Ann Arbor.

Gray, H., W. Schucany, and T. Watkins (1975). On the generalized jackknife and its relation to statistical differentials. <u>Biometrika</u>, <u>62</u>, 637-642.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. <u>Journal of the American Statistical Association</u>, <u>69</u>, 383-93.

Hinkley, D. V. (1978). Improving the jackknife with special reference to correlation estimation. <u>Biometrika</u>, <u>65</u>, 13-21.

Hislop, D. and S. Lemeshow (1977). Evaluation of balanced half-sample and jackknife estimates of combined ratio estimates for non-normally distributed populations. <u>Proceedings of the Social Statistics Section</u>, 843-847. American Statistical Association, Washington, D.C.

Jaeckel, L. (1972). The infinitesimal jackknife. Technical memorandum MM 72-1215-11, Bell Laboratories, Murray Hill, New Jersey.

Kish, L. and M. R. Frankel (1970). Balanced repeated replications for standard errors. <u>Journal of the American Statistical Association</u>, <u>65</u>, 1071-1094.

Krewski, D. and J.N.K. Rao (1978). Inference from stratified samples I: large sample properties of the linearization, jackknife and balanced repeated replication methods. Technical Report No. 155, Carleton University Mathematical Series.

McCarthy, P. J. (1966). Replication: an approach to the analysis of data from complex surveys. <u>Vital and Health Statistics</u>, <u>Series 2</u>, No. 14. National Center for Health Statistics, Washington, D.C.

McCarthy, P. J. (1969a). Pseudoreplication: further evaluation and application of the balanced half-sample technique. <u>Vital and Health Statistics</u>, <u>Series 2</u>, No. 31. National Center for Health Statistics, Washington, D.C.

McCarthy, P. J. (1969b). Pseudoreplication: half samples. <u>Review of the International Statistical Institute</u>, <u>37</u>, 239-264.

Mellor, R. W. (1973). <u>Subsample Replication Variance Estimators</u>. Doctoral dissertation, Harvard University.

Plackett, R. L. and P. J. Burman (1946). The design of optimum multifactorial experiments. Biometrika, 33, 305-25.

Rao, C. R. (1973). Introduction to Linear Statistical Inference, second edition. Wiley, New York.

Reeds, J. A. (1976). On the Definition of von Mises Functionals. Doctoral dissertation, Harvard University.

Tepping, B. J. (1968). The estimation of variance in complex surveys. Proceedings of the Social Statistics Section, 11-18. American Statistical Association, Washington, D.C.

von Mises, R. (1947). On the asymptotic distribution of differentiable statistical functions. Annals of Mathematical Statistics, 18, 309-348.

Walsh, J. E. (1947). Concerning the effect of intraclass correlation on certain significance tests. Annals of Mathematical Statistics, 18, 88-96.

Appendix

The second order von Mises approximation for $\hat{\theta}$ is used to evaluate $MSE(\hat{\theta})$ and $E[\hat{V}_J(\hat{\theta})]$. Although equality signs will be used throughout, the reader is reminded that all results are approximations. Let $\{y_{ij}: i=1, \ldots, L; j=1,2\}$ be realizations of general independent random variables $Y_{ij} \sim F_i(y)$. Then

$$\hat{\theta}-\theta = \frac{1}{2L} \sum_{i=1}^{L} \sum_{j=1}^{2} t_1(y_{ij}) + \frac{1}{2(2L)^2} \left[\sum_{i,j}^{L} \sum_{k,\ell}^{2} t_2(y_{ik}, y_{j\ell}) \right] . \qquad (A.1)$$

The derivatives t_1 and t_2 are defined by (4.2) and (4.3); they satisfy

$$E_{\overline{F}} [t_1(Y)] = 0 \qquad (A.2)$$

$$E_{\overline{F}} [t_2(a,Y)] = 0 \qquad (A.3)$$

$$t_2(x,y) = t_2(y,x) \qquad (A.4)$$

with

$$\overline{F}(y) = \frac{1}{L} \sum_{i=1}^{L} F_i(y) \quad . \qquad (A.5)$$

We now introduce the necessary notation:

$$E_1(i) = E_{F_i} [t_1(Y)]$$

$$E_2(i) = E_{F_i} [t_2(Y,Y)]$$

$$E_3(i,j) = E_{F_i,F_j} [t_2(Y,Z)] = E_3(j,i)$$

$$V_1(i) = V_{F_i} [t_1(Y)]$$

$$V_2(i) = V_{F_i} [t_2(Y,Y)]$$

$$V_3(i,j) = V_{F_i,F_j} [t_2(Y,Z)] = V_3(j,i)$$

$$C_1(i) = \mathrm{Cov}_{F_i}[t_1(Y), t_2(Y,Y)]$$

$$C_2(i,j) = \mathrm{Cov}_{F_i,F_j}[t_1(Y),\ t_2(Y,Z)]$$

$$C_3(i,j) = \mathrm{Cov}_{F_i,F_j}[t_2(Y,Y),\ t_2(Y,Z)]$$

$$C_4(i;j,k) = \mathrm{Cov}_{F_i,F_j,F_k}[t_2(Y,Z), t_2(Y,X)] = C_4(i;\ k,j) \quad .$$

As a consequence of (A.2) and (A.3) we have the following relationships:

$$\left.\begin{array}{ll} \sum_{i=1}^{L} E_1(i) = 0 & \sum_{j=1}^{L} E_3(i,j) = 0 \\ \sum_{j=1}^{L} C_2(i,j) = \sum_{j=1}^{L} C_3(i,j) = 0 & \\ \sum_{j=1}^{L} C_4(i;\ j,k) = \sum_{k=1}^{L} C_4(i;\ j,k) = 0 \quad . & \end{array}\right\} \qquad (A.6)$$

For convenience we reexpress (A.1) as

$$\hat{\theta}-\theta = \frac{1}{2L}\sum_{i=1}^{L}\sum_{j=1}^{2} t_1(y_{ij}) + \frac{1}{2(2L)^2}\sum_{i=1}^{L} S(i,i) + \frac{1}{(2L)^2}\sum_{i<k}^{L}\sum^{L} S(i,k), \qquad (A.7)$$

$$= \quad T_1 \quad + \quad T_2 \quad + \quad T_3$$

where $S(i,i) = t_2(y_{i1},y_{i1}) + t_2(y_{i2},y_{i2}) + 2t_2(y_{i1},y_{i2})$,

$$S(i,k) = t_2(y_{i1},y_{k1}) + t_2(y_{i1},y_{k2}) + t_2(y_{i2},y_{k1}) + t_2(y_{i2},y_{k2}) \ ,$$

and the definitions of T_1, T_2, and T_3 are evident.

It is now straight-forward but tedious to evaluate the variances and covariances of T_1, T_2, and T_3. Doing this, dropping terms of $O(L^{-3})$, and using (A.6) gives the desired result:

$$V(\hat{\theta}) = \frac{1}{2L^2}\sum_{i=1}^{L} V_1(i) + \frac{1}{8L^4}\sum_{i\neq j}^{L}\sum^{L} V_3(i,j) - \frac{1}{4L^4}\sum_{i\neq j}^{L}\sum^{L} C_4(i;\ j,j)$$

$$+ \frac{1}{4L^3}\sum_{i=1}^{L} C_1(i) - \frac{1}{2L^3}\sum_{i=1}^{L} C_2(i,i) \quad . \qquad (A.8)$$

Taking the expectation of (A.1) and simplifying with (A.6), we find

$$E[\hat{\theta}-\theta] = \frac{1}{(2L)^2} \sum_{i=1}^{L} [E_2(i) - E_3(i,i)] \quad , \tag{A.9}$$

and $$\text{MSE}(\hat{\theta}) = V(\hat{\theta}) + E^2[\hat{\theta}-\theta] \ . \tag{A.10}$$

We now find

$$E[\hat{V}_J(\hat{\theta})] = E\left[\frac{1}{2} \sum_{i=1}^{L} \sum_{j=1}^{2} (\hat{\theta}^J_{ij} - \hat{\theta})^2\right]$$

by using the second order approximation for both $\hat{\theta}^J_{ij}$ and $\hat{\theta}$. Remembering that $\hat{\theta}^J_{ij}$ is the pseudoestimate of θ formed by deleting y_{ij} and including $y_{ij'}$ twice, we see that

$$\begin{aligned}\hat{\theta}^J_{p1} - \theta = {} & \frac{1}{2L}\left\{\sum_{i\neq p}^{L} [t_1(y_{i1}) + t_1(y_{i2})] + 2t_1(y_{p2})\right\} \\ & + \frac{1}{2(2L)^2}\left\{\sum_{i\neq p}^{L} S(i,i) + 4t_2(y_{p2},y_{p2})\right\} \\ & + \frac{1}{2(2L)^2}\left\{\sum_{i\neq k\neq p}\sum S(i,k) + 2\sum_{i\neq p} [2t_2(y_{i1},y_{p2}) + 2t_2(y_{i2},y_{p2})]\right\} .\end{aligned} \tag{A.11}$$

Subtracting (A.7) from (A.11) gives

$$\hat{\theta}^J_{p1} - \hat{\theta} = \frac{1}{2L} D_1(p) + \frac{1}{2(2L)^2} D_2(p,1) + \frac{1}{(2L)^2}\left[\sum_{i\neq p}^{L} R(p,i)\right] \ . \tag{A.12}$$

We also obtain

$$\hat{\theta}^J_{p2} - \hat{\theta} = \frac{-1}{2L} D_1(p) + \frac{1}{2(2L)^2} D_2(p,2) - \frac{1}{(2L)^2}\left[\sum_{i\neq p}^{L} R(p,i)\right]$$

where

$$D_1(p) = t_1(y_{p2}) - t_1(y_{p1})$$

$$D_2(p,1) = 3t_2(y_{p2},y_{p2}) - t_2(y_{p1},y_{p1}) - 2t_2(y_{p1},y_{p2})$$

$$D_2(p,2) = 3t_2(y_{p1},y_{p1}) - t_2(y_{p2},y_{p2}) - 2t_2(y_{p1},y_{p2})$$

$$R(p,i) = t_2(y_{i1},y_{p2}) + t_2(y_{i2},y_{p2}) - t_2(y_{i1},y_{p1}) - t_2(y_{i2},y_{p1}) \ .$$

From (A.12),

$$\sum_{p=1}^{L}\sum_{j=1}^{2}(\hat{\theta}^{J}_{pj}-\hat{\theta})^2 = \frac{1}{2L^2}\sum_{p=1}^{L} D_1^{\,2}(p) + \frac{1}{4(2L)^4}\sum_{p=1}^{L}[D_2^{\,2}(p,1) + D_2^{\,2}(p,2)]$$

$$+ \frac{2}{(2L)^4}\sum_{p=1}^{L}\left[\sum_{i\neq p}^{L} R(p,i)\right]^2 + \frac{1}{(2L)^3}\sum_{p=1}^{L} D_1(p)\ [D_2(p,1) - D_2(p.2)]$$

$$+ \frac{4}{(2L)^3}\sum_{p=1}^{L} D_1(p)\left[\sum_{i\neq p}^{L} R(p,i)\right] + \frac{1}{(2L)^4}\sum_{p=1}^{L}\left[\sum_{i\neq p}^{L} R(p,i)\right]\left[D_2(p,1) - D_2(p,2)\right] . \tag{A.13}$$

Finding $\frac{1}{2}$ E[A.13] = $E[\hat{V}_J(\hat{\theta})]$ is simplified by recalling

$$E[X^2] = V(X) + E^2(X)$$

and

$$E[XY] = Cov[X,Y] + E(X)E(Y) \quad .$$

The result, after dropping $O(L^{-3})$ terms, is

$$E[\hat{V}_J(\hat{\theta})] = \frac{1}{2L^2}\sum_{i=1}^{L} V_1(i) + \frac{1}{2L^4}\sum_{i<j}^{L}\sum^{L} V_3(i,j) - \frac{1}{2L^4}\sum_{i\neq j}^{L}\sum^{L} C_4(i;\,j,j)$$

$$+ \frac{1}{2L^3}\sum_{i=1}^{L} C_2(i) - \frac{1}{L^3}\sum_{i=1}^{L} C_2(i,i) \tag{A.14}$$

Comparing (A.8) and (A.14) leads to the definition of R as

$$V(\hat{\theta}) = \frac{1}{2L^2}\sum_{i=1}^{L} V_1(i) + R.$$

$$E[\hat{V}_J(\hat{\theta})] = \frac{1}{2L^2}\sum_{i=1}^{L} V_1(i) + 2R \quad .$$

and the second order bias in the $\hat{V}_J(\hat{\theta})$ is

$$E[\hat{V}_J(\hat{\theta})] - MSE(\hat{\theta}) = R - E^2[\hat{\theta}-\theta].$$